

Review for Final

- Review Session this Thursday from 3-5pm in SEO 512
- Office Hours this week and next are Tuesdays 1-3pm in the MLC (SEO 430). Next week's office hours will be a review session if enough people show up.
- *Bring questions to the review session to make the most out of it.*

Sick Trees

- A researcher is studying two forests for the prevalence of Hemlock Fever in the Hemlocks. She randomly samples hemlock trees in both forests. In forest 1, 382 out of 1035 hemlocks are infected. In forest 2, 172 out of 422 are infected.
- Her hunch is that the disease is more prevalent in Forest 2. What should the hypothesis test be? (Use $p_2 - p_1$ as a statistic and a significance level of .05)

Sick Trees (cont)

- $H_0: p_2 - p_1 = 0$ vs $H_A: p_2 - p_1 > 0$
- Observed $p_1 = 382/1035 = .3691$
- Observed $p_2 = 172/422 = .4075$
- Calculate z under the assumption of H_0 :
 - Under H_0 ,
$$SE = \sqrt{(p_{\text{pooled}} q_{\text{pooled}} / n_1 + p_{\text{pooled}} q_{\text{pooled}} / n_2)}$$
 - $p_{\text{pooled}} = (382 + 172) / (1035 + 422) = .3802$
 - $SE = \sqrt{(.3802 * .6198 / 1035 + .3802 * .6198 / 422)} = 0.02803$

Sick Trees (cont)

- $p_2 - p_1 = .4075 - .3691 = .0384$
- $z = .0384 / .02803 = 1.3699$
- Because $H_A: p_2 - p_1 > 0$, p-value is
 $\text{Prob}(Z > 1.3699) = \text{normalcdf}(1.3699, 6) = .08536$
- This is NOT lower than our significance level, so we do not reject the null hypothesis; we do not conclude that forest 2 is “sicker” than forest 1

Confidence Interval

- If we had concluded forest 2 is sicker, the 95% confidence interval for the true $p_2 - p_1$ is:
- $.0384 \pm 1.96 * SE(p_2 - p_1)$
- For the confidence interval,
 $SE = \sqrt{(.3691 * .6309 / 1035 + .4075 * .5925 / 422)} = 0.02823$
- Confidence interval is $-.0169$ to $.09373$